

ABSTRACT OF THE DISCLOSURE

A computer software architecture to automatically optimize the throughput of the data extraction/transformation/loading (ETL) process in data warehousing applications. This architecture has a componentized aspect and a pipeline-based aspect. The componentized aspect refers to the fact that every transformation used in this architecture is built up with transformation components selected from an extensible set of transformation components. Besides simplifying source code maintenance and adjustment for the data warehouse users, these transformation components also provide these users the building blocks to effectively construct pertinent and functionally sophisticated transformations in a pipelined manner. Within a pipeline, each transformation component automatically stages or streams its data to optimize ETL throughput. Furthermore, each transformation either pushes data to another transformation component, pulls data from another transformation component, or performs a push/pull operation on the data. Thereby, the pipelining; staging/streaming; and pushing/pulling features of the transformation components effectively optimizes the throughput of the ETL process.

20

CONFIDENTIAL